

INTERACTOR: Agentic RL oriented Iterative Creation for Ad Description Generation in Sponsored Search

Penghui Wei, Jiayu Wu, Chao Ye, Zhi Guo, Shuanglong Li and Lin Liu

Search Ads, Baidu Inc.

{weipenghui,wujiayu01,yechao01,guozhi,lishuanglong,liulin03}@baidu.com

Abstract

This paper focuses on automatically generating informative ad descriptions in sponsored search. Unlike ad titles which are usually optimized to attract user click feedbacks, ad descriptions have a longer text span and possess the potential of incorporating world knowledge to address user search intents while presenting the fine-grained selling points of the ads. We propose INTERACTOR, a multi-turn iterative creation framework optimized with agentic RL for ad description generation. The generation model acts as a policy that interacts with a customized environment consisting of multiple generative reward models. Given initial generations by the policy, the customized Gen-RMs evaluate qualities including knowledge capacity and landing page consistency, providing both binary signals and detailed feedbacks. The policy then iteratively refines the descriptions based on such feedbacks to ensure continuous improvement. Experiments show that it significantly outperforms state-of-the-art ad text generation approaches in generating knowledge-rich and faithful ad descriptions. Since late May 2026, it has been deployed online in a leading search ads system, where the framework serves over 140k advertisers, contributing to both ad revenue and user experience.

1 Introduction

Search advertising plays a crucial role to bridge user intents and business services, which brings substantial revenue to search engine platforms. Advertisers usually put a lot of effort into designing ad creatives (e.g., title, description and image) to clearly show the advantages of their services and attract potential audiences. With the rise of generative models and AI generated content (AIGC) techniques, extensive research proposed to automatically generate ad texts and images to improve click-through rate (CTR), achieving better delivery performance compared to the creatives provided by

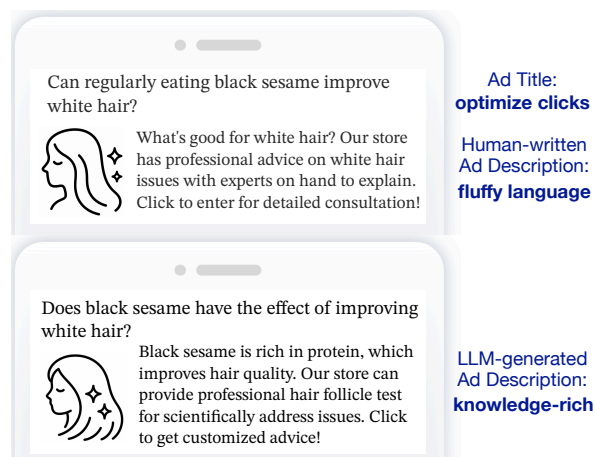


Figure 1: An example of ad description with query ‘Can black sesame turn white hair black?’. Compared to title, it has a longer span and should be more informative.

advertisers. Most work on ad text generation has focused on ad titles, while ad descriptions – which also occupy a large visual space and allow for more informative contents – remain underexplored. An ideal ad description should exploit the superiority of longer text span, integrating both 1) relevant and appropriate world knowledge that matches user’s search intent, and 2) fine-grained selling points mined from landing pages to faithfully represent the advantages of services or products.

Ad description generation is an open-ended task with no ground-truth reference to imitate. Reinforcement learning, which alternates between collecting rollouts for exploration and optimizing the policy with rewards for exploitation, is therefore a natural solution, especially in online advertising where user preferences change dynamically. The state-of-the-art approaches for ad text generation (Chen et al., 2025; Wang et al., 2026) learn to generate titles achieving better quality and higher CTR via preference optimization or RLHF, with implicit or explicit reward modeling to guide LLM training via scalar signals. However, the ad de-

scription generation task cannot directly adopt the above approaches heavily relying on scalar rewards during optimization. Compared to ad titles, ad descriptions are relatively long and must satisfy more quality constraints, thus the generated results inevitably suffer from various subtle inaccuracies. Current approaches struggle to produce detailed guidance information for revising incorrect generations, which results in inefficient exploration due to low-quality rollouts.

To tackle the above challenges, in this paper we propose INTERACTOR, an agentic RL oriented iterative creation framework for informative ad description generation.¹ Compared to generating the final description in a single pass optimized by scalar rewards, the INTERACTOR framework incorporates a multi-turn iteration process, in which the LLM to be optimized for description generation is regarded as an agent that receives detailed feedback from customized generative reward models to refine previous generations for improvement.

Specifically, the LLM policy interacts with a customized environment consisting of GenRMs responsible for evaluating multi-dimensional qualities such as knowledge capacity and landing page consistency. After the initial descriptions generated by the policy, the GenRMs discriminate the generations with binary signals and also provide detailed reasoning feedbacks (e.g., the reason why the initial generations miss appropriate world knowledge related to user intent, or pointing out the unfaithful contents inconsistent with landing pages), based on the predefined rubrics summarized from business requirements. If the criteria are not met, the policy then iteratively refines the descriptions via incorporating such feedbacks into context to form a multi-turn iterative creation process. The policy is optimized via agentic RL with multi-dimensional rewards to ensure that the quality of subsequent generations improves continuously.

Since late May 2026, the INTERACTOR framework has been deployed online in a leading search ads system, contributing to both ad revenue and user experience. The main contributions are:

- We identify the necessity to shift from CTR-driven title generation to informative description generation in industrial practice, highlighting the importance of explicitly modeling world knowledge to satisfy user intent.

- We propose a multi-turn iteration framework INTERACTOR for ad description generation, in which the LLM policy is optimized via agentic RL with customized generative rewards to achieve continuous improvement on knowledge capacity and landing page consistency.
- Experiments on industrial datasets show that the INTERACTOR framework significantly outperforms state-of-the-art single-turn approaches in generating knowledge-rich and faithful ad descriptions. Online A/B test verifies its effectiveness in improving ad revenue and user experience.

2 Problem Definition

We introduce the task definition of ad description generation in sponsored search. Given a user search query x_{user} and an ad’s landing page information x_{ad} , the task’s objective is to learn a generation policy $\pi_{\theta}(y \mid x_{\text{user}}, x_{\text{ad}})$ that outputs an informative description y satisfying quality requirements as well as achieving higher CTR, taking both user experience and advertiser value into account.²

The quality is evaluated by two dimensions. 1) Knowledge capacity: as a long-form text (compared to ad title), the ad description y is expected to contain world knowledge that is relevant to user search intent x_{user} for improving informativeness. 2) Landing page consistency: the ad description y ’s detailed content that represents the advantages of services or products must be strictly entailed by the landing page information x_{ad} , without any unfaithful exaggerations that mislead users.

3 Methodology: The INTERACTOR Framework

Unlike current ad text generation approaches which generate the results in a single-turn process, the INTERACTOR framework formulates the description generation process as *multi-turn iteration* optimized with agentic RL, where the LLM policy interacts with a customized environment that provides detailed feedbacks about content quality for continuous improvement. Our framework can refine incorrect expressions with fine-grained feedbacks, which is suitable for open-ended, long-form ad description generation.

²In practice, to provide more information as input, we also use the keyword (retrieved based on user query or recommended by the system) and ad metadata (e.g., advertiser name and trade name) to supplement x_{user} and x_{ad} , respectively.

¹INTERACTOR: Agentic Reinforcement Learning for Ad Description Generation.

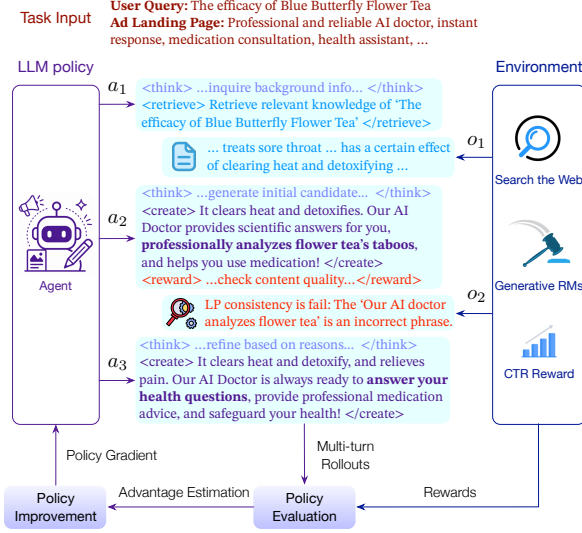


Figure 2: The overall workflow of the INTERACTOR framework for generating informative ad descriptions.

3.1 Formulation: Multi-Turn Iteration

Figure 2 shows the overview of INTERACTOR, where an LLM policy interacts with the customized environment to produce the final generation.

Formally, the initial context consists of customized task prompt x_{task} and inputs $\{x_{\text{user}}, x_{\text{ad}}\}$. The task prompt x_{task} strictly structures the LLM policy’s output format of each turn into two parts:

1. A thinking process (enclosed in special tokens `<think>` `</think>`) that understands context including task inputs and environment’s observations from previous turns;
2. A response that reflects a real **action**, which may *request the search engine* (enclosed in `<retrieve>` `</retrieve>`) for injecting relevant world knowledge, *generate an ad description* based on all turns’ contexts (enclosed in `<create>` `</create>`), and *request reward models* (enclosed in `<reward>` `</reward>`) for fine-grained evaluation and refinement.

At the t -th turn, the LLM policy takes an **action** a_t (thinking and response) based on the context $c_t = ([x_{\text{task}}; x_{\text{user}}; x_{\text{ad}}], a_1, o_1, \dots, a_{t-1}, o_{t-1})$,³ then receives an **observation** o_t (e.g., retrieved knowledge and reasoning information of rewards) from the environment and also a **reward** r_t (multi-dimensional evaluations of qualities and CTR) if the action is generation. The objective is to optimize the policy $\pi_{\theta}(y \mid x_{\text{user}}, x_{\text{ad}})$ guided by the reward signals with policy gradient.

³For the first turn, the context $c_1 = ([x_{\text{task}}; x_{\text{user}}; x_{\text{ad}}])$.

3.2 The Customized Environment for Ad Description Generation

To perform multi-turn iteration, we design a customized environment that provides observations to the learned policy through generative reward models and a knowledge retrieval engine.

3.2.1 Rewards: Hybrid of Rubric-based Generative RMs and Human Feedback

Rubric-based GenRMs for Quality. To evaluate the qualities and also guide further refinement, we build two generative reward models for knowledge capacity and landing page consistency – the main dimensions of content quality – based on predefined rubrics to transform open-ended evaluation to verifiable signals (Kim et al., 2024; Gunjal et al., 2026). Unlike current state-of-the-art ad text generation approaches which only guide RLHF via scalar rewards, our GenRMs produce both binary signals and fine-grained reasoning information.

Beyond scalar rewards for exploitation, the detailed feedbacks facilitate next turn’s refinement of ad descriptions, directly contributing to high-quality rollouts for efficient exploration. As shown in Figure 2, for landing page consistency, the GenRM points out unfaithful phrases that mislead users. For knowledge capacity, it can return the reason why the initial description misses appropriate world knowledge that matches user intent.

Specifically, for each quality dimension, we first summarize the rubrics consisting of multiple verifiable rules according to business requirements, and then prompt LLMs with few-shot in-context learning to obtain 1) binary result of the reward, if and only if all rubrics are satisfied then the quality reward is 1, and 2) fine-grained reasoning information that explains the improvement directions for a given description. The predefined rubrics, prompts and other details of the GenRMs can be found in Appendix B.1.

CTR Estimation with Human Feedback. To evaluate the attractiveness of generated descriptions, we build a CTR estimation model trained on the online logs of user feedbacks in our search ads system. It serves as a reward that produces a real number $r_t^{\text{CTR}} \in (0, 1)$ as the predicted CTR given both task inputs and a generated description.

To capture the dynamics of user preference, the CTR estimation model is updated incrementally in a daily fashion. For each day we continually train the model via loading previous day’s checkpoint and optimizing parameters with fresh data.

3.2.2 Observations: Reasoning from GenRMs and Knowledge from Search Engine

The description generation model acts as an LLM policy $\pi_\theta(\cdot | c_t)$ that interacts with the environment to receive observations to augment contexts for further actions. The text observation o_t is typically the reasoning information of GenRMs for quality evaluation and refinement, or the retrieved knowledge from the search engine for generating contents that satisfy user search intent.

Reasoning from Quality GenRMs. For a turn with the context c_t , if the policy’s action a_t is to generate a description, the GenRMs will be requested to provide both scalar signals and reasoning information with fine-grained feedbacks about incorrect phrases in the description. Such reasoning information is used as observation o_t to augment next turn’s context c_{t+1} , guiding the policy to inject missing world knowledge or modify unfaithful contents that are inconsistent with the landing page.

Retrieved Knowledge from Search Engine. The policy’s action a_t may request the search engine (triggered by generating content in `<retrieve>` `</retrieve>`) to retrieve relevant knowledge based on the search query x_{user} (usually at the first turn), which helps the generated ad description contain appropriate world knowledge that matches user search intent. Such retrieved knowledge is regarded as observation o_t to augment next turn’s context c_{t+1} , which supplements the landing page information to enhance the task input.⁴

Specifically, we use our in-house search engine, which indexes a massive corpus of encyclopedic web pages and thus is highly well-suited for acquiring world knowledge. Given the search query x_{user} , the document of the top-1 retrieved web page is used as the knowledge to augment context.

3.3 Agentic RL for Optimizing Description Generation Policy

For each turn t , the description generation model takes an action $a_t \sim \pi_\theta(\cdot | c_t)$ based on the context c_t containing task inputs, previous actions and observations. It is expected that at subsequent turns the policy can generate descriptions with better quality and higher CTR given the fine-grained reasoning from the environment.

⁴Although the base LLM of the generation model stores world knowledge implicitly in the parameters, in practice we found that requesting our in-house search engine to explicitly enhance the context can achieve better generation results.

3.3.1 Multi-Turn Rollouts

Based on the iterative creation nature, each episode constitutes a multi-turn rollout sequence $\tau = ([x_{\text{task}}; x_{\text{user}}; x_{\text{ad}}], a_1, o_1, r_1, \dots, a_T, o_T, r_T)$, where each action’s token sequence structure contains a thinking sequence and a response sequence (Yao et al., 2023), and the *final generated description* of the LLM policy is extracted from the response part of the last turn’s action a_T .

Specifically, at subsequent turns, the LLM policy generates descriptions that refine incorrect phrases with the help of the reasoning information from the environment. As shown in Figure 2, the subsequent turn successfully refines the initial description via modifying unfaithful phrase and also injecting knowledge guided by the detailed feedback, achieving improvement via iterative creation.

Here the reward signal r_t is the weighted sum of quality rewards and predicted CTR:

$$r_t = \sum_{* \in \{\text{KC}, \text{LP}, \text{CTR}\}} w^* \cdot r_t^* \quad (1)$$

where $r_t^{\text{KC}} / r_t^{\text{LP}}$ denote the binary signal from the GenRM of knowledge capacity / landing page consistency, and r_t^{CTR} denotes the predicted CTR, with their factors w^* .

3.3.2 Optimization

We use group sequence policy optimization (GSPO) (Zheng et al., 2025) to learn the LLM policy, which is critic-free (Shao et al., 2024) and defines a sequence-level importance ratio to match the rewarding granularity. Formally, we maximize:

$$\mathbb{E}_{\{\tau^{(i)}\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G \left(\frac{\pi_\theta(\tau^{(i)})}{\pi_{\theta_{\text{old}}}(\tau^{(i)})} \right)^{\frac{1}{|\tau^{(i)}|}} A_{\text{norm}}^{(i)} \right] \quad (2)$$

Here we omit detailed clip operation and KL term for brevity. The advantage $A_{\text{norm}}^{(i)}$ of each rollout sequence $\tau^{(i)}$ per prompt is computed using the last turn’s reward $r_T^{(i)}$ via group-based estimation of G rollouts and global normalization in each training step’s data $\mathcal{D}_{\text{step}}$ (Hu et al., 2025):

$$A_{\text{norm}}^{(i)} = \text{Norm} \left(\text{Norm}(r_T^{(i)} | i \in [1, G]) | \mathcal{D}_{\text{step}} \right). \quad (3)$$

Note that in a rollout sequence τ , a token may be generated by the LLM policy (i.e., a_t), provided by the environment (i.e., o_t), or come from the task input (i.e., $[x_{\text{task}}; x_{\text{user}}; x_{\text{ad}}]$). Thus during policy gradient computation, we only reserve action tokens while masking the others.

4 Experiments

4.1 Experimental Setup

Dataset Construction To our knowledge, no publicly available benchmark exists for ad description generation,⁵ thus we verify the effectiveness of INTERACTOR on a large-scale proprietary dataset collected from our search ads system, operating as a core component of a leading search engine in Chinese. We sample the impression log of search ads to obtain 1 million user-ad pairs as the training set, and hold-out 5k / 10k pairs as the validation / test set where the impression time is strictly later than training data.

Evaluation Metrics We thoroughly evaluate the generations with five dimensions: 1) Basic quality: measures whether the content contains basic errors such as wrongly-written characters / punctuations and repetitive phrases. 2) Informativeness: indicates whether the description contains knowledge relevant to user query. 3) Faithfulness: indicates the landing page consistency. 4) Attractiveness: reflects whether the user intends to click on. 5) Diversity: measures the content similarity across multiple candidates. More details of the evaluation process are listed in Appendix B.2.

Competitors We use the state-of-the-art approaches of ad text generation, as well as our best-performing production model as baselines:

- CTOP (Chen et al., 2025) and DIVER (Wang et al., 2026): the SOTA approaches for ad title generation via preference optimization and RLHF respectively to jointly optimize quality and CTR. We directly adopt them to perform description generation task.
- SINGLE-TURN RL: the best-performing production policy of our system that incorporates the same knowledge retrieval and reward design as INTERACTOR. It is optimized via single-turn RL and thus generates the final description without iteration.

Implementation Details For fair comparison, we unify the base model as Qwen3-30B-A3B (Qwen, 2025), and employ GSPO for all RL-based competitors. We choose Slime⁶ and SGLang (Zheng et al., 2024) as the post-training and inference framework, respectively. Appendix B.3 lists more details.

⁵Current open-source benchmark of ad text generation focuses on ad title (Mita et al., 2024).

⁶<https://github.com/THUDM/slime>

Approach	Basic.	Info.	Faith.	Attract.	Div.
CTOP	0.989	0.274	0.634	7.8%	0.662
DIVER	0.985	0.359	0.724	11.0%	0.684
SINGLE-TURN	0.986	0.693	0.726	12.3%	0.715
INTERACTOR	0.986	0.715	0.872	12.4%	0.732

Table 1: Main results of basic quality, informativeness, faithfulness, attractiveness and diversity.

4.2 Main Results

Table 1 reports the evaluation results of all compared approaches. For each input we generate ten candidates, and report Average@10 as the results.

The superiority of RL-based approaches compared to CTOP shows that learning from pre-collected static data with preference optimization is sub-optimal in open-ended ad description generation task, especially for the attractiveness metric which is measured by predicted CTR that reflects user’s dynamic interests. For the informativeness metric, both INTERACTOR and our production model SINGLE-TURN RL outperform other competitors by a large margin, demonstrating that explicitly incorporating world knowledge is crucial for generating high-quality descriptions that satisfy user intent, which reveals the essential difference between ad description and ad title.

The INTERACTOR framework significantly outperforms other competitors on the whole, especially for the faithfulness metric which is crucial but difficult to optimize for ad text. The results verify the effectiveness of the multi-turn iterative creation for optimizing landing page consistency with the help of fine-grained reasoning information, and also indicate that the optimization paradigm of agentic RL possesses the ability of improving various metrics simultaneously.

4.3 Discussion

Effects of Multi-turn Iteration During the inference stage of INTERACTOR, we also extract the generated description from the penultimate turn and report the metrics in Table 2. We can see that the last turn beats the penultimate turn on most metrics, demonstrating that the iterative creation process brings continuous improvement on description quality. Moreover, the penultimate turn still outperforms SINGLE-TURN RL, which shows that the agentic RL framework indeed learns to generate better descriptions rather than merely suppressing previous turns to inflate the final improvement.

Approach	Basic.	Info.	Faith.	Attract.
INTERACTOR	0.986	0.715	0.872	12.4%
penultimate	0.982	0.694	0.798	12.5%
w/o reasoning	0.985	0.636	0.706	12.3%
smaller GenRMs	0.985	0.692	0.861	12.3%

Table 2: Ablation studies of multi-turn iteration, reasoning, and GenRM choice for INTERACTOR.

Ablation of Reasoning for Rollouts We further verify the effectiveness of the detailed reasoning information from the environment. Specifically, we train an LLM policy which omits the guide of reasoning information from the GenRMs and performs multi-turn rollouts based on the binary signals only. As the results of “w/o reasoning” in Table 2, the performance drops dramatically, which reveals that the decisive factor of INTERACTOR’s iterative creation process is the fine-grained feedbacks of the GenRMs and indicates the effectiveness of the overall multi-turn framework.

Sensitivity to Model Choice of GenRMs Based on the above results, we further analyze how sensitive the performance is to the choice of GenRM backbone. Specifically, we replace the GenRMs from the Qwen3-30B-A3B to the smaller-sized Qwen3-8B (Qwen, 2025) with the same prompts. The results of “smaller GenRMs” in Table 2 verify that our framework is not very sensitive to the choice of GenRMs, however stronger GenRMs can further improve the performance of generated descriptions on multiple metrics.

4.4 Case Analysis

In Table 3 we qualitatively show a generated case. We observe that the subsequent turn successfully refines the initial description via modifying unfaithful phrase guided by the detailed feedback from the customized GenRMs and injecting knowledge relevant to user query, achieving a significant quality improvement via reasoning-guided iterative creation. This also illustrates that the phenomenon where the penultimate turn achieves a slightly higher CTR than the final turn is not a flaw, but the intended behavior of our policy: in earlier turn the model may generate attractive but unsupported claims, then the later turn utilizes fine-grained feedbacks to explicitly suppress these unfaithful elements, replacing them with factual ones.

Input	User Query: 电钢琴型号推荐 (Recommended electric piano models) Landing Page: 家用电钢琴各种型号,雅马哈/美得理厂家直销原装联保,便携式可折叠/重锤,考级/初学/儿童/成人,送上门... (Various home electric pianos, Yamaha/Medeli factory direct sales, original packaging and warranty, portable and foldable/heavy hammer, examination/beginner/children/adult, door-to-door delivery...)
Action a_1	检索“电钢琴型号推荐”相关知识 (Retrieve relevant knowledge of ‘Recommended electric piano models’)
Observation o_1	...初学者买电钢琴,首先考虑口碑和声誉,选购时需亲自体验手感,音质也是不能忽略的因素... (...For beginners of purchasing an electric piano, it is important to first consider its reputation and word-of-mouth, essential to personally experience the feel of the keys. The sound quality is also a factor that cannot be ignored...)
Action a_2	电钢琴型号推荐!雅马哈等品牌热销,重锤键盘专业级手感,初学考级皆适用,原装正品,快速包邮,便携折叠款可选,限时优惠! (Recommended electric piano models! Popular brands such as Yamaha, featuring heavy-hammer keyboards for professional-level touch, suitable for beginners and grade exams. Genuine products, free shipping via express delivery, and portable folding models available. Limited-time discount!)
Observation o_2	知识化程度不通过: 检索到电钢琴选购知识,但未体现在描述中 落地页一致性不通过: 卖点“包邮”的原意及类似含义未出现在广告信息中。 (Knowledge capacity is fail: Knowledge about electric piano was retrieved, but it was not integrated. LP consistency is fail: The selling point of ‘free shipping’ mentioned in the candidate is not found in the ad information with a similar meaning.)
Action a_3	新手选购电钢琴需兼顾键盘质量、手感及品牌口碑!热销雅马哈等品牌,重锤键盘模拟真实手感,适合初学考级,原装联保,快速上楼,便携折叠设计,限时优惠! (For beginners of electric pianos, it’s essential to consider keyboard quality, feel, and brand reputation! Popular brands like Yamaha offer heavy-hammer keyboards simulating real touch, ideal for beginners and grade exams. With original warranty, door-to-door delivery, portable folding design, and limited-time discounts!)

Table 3: Case analysis. We omit the sequence enclosed in <think> </think> for brevity. Contents in the parentheses are the results translated to English.

4.5 Online A/B Test and Human Evaluation

We conduct online A/B test in our search ads system, and the experiment group covers over **140k advertisers** and accumulated over **60 million ad impressions**. Results verify that INTERACTOR contributes to both ad revenue (**+0.74%**) and user experience (top-grade rate **+7pp** of content quality in human evaluation). Since late May 2026, it has been deployed online to serve the main traffic.

Deployment The descriptions from INTERACTOR are generated with offline batching inference and cached in a KV database for rapid retrieval during real-time auction. Overall, the dynamic creative module in our system consists of a generation stage and a ranking stage. After the ad pre-ranking module, the creative generation stage assembles candidates via 1) retrieving advertiser-uploaded assets and offline-generated titles, descriptions and images from KV databases; and 2) generating titles in real time (Dong et al., 2026), where this real-time

branch is selectively deployed only for high-value traffic. Subsequently, the creative ranking stage performs personalized scoring at the creative combination level, selecting the top-1 combination to downstream ad ranking module.⁷ Because multiple offline branches exist for each creative type, a single ad can yield thousands of potential combinations. During experiment, the candidate size allocated to the INTERACTOR branch is two per ad, and concurrently we deprecate an old branch.

Online A/B Test For online A/B test, the randomization unit is at the search page view (PV) level. Our experiment achieved **+0.74%** lift on ad revenue (where the 95% confidence interval is [0.13%,1.40%] with p-value 0.02). Guardrail metrics, such as the App’s daily active users (DAU), search PV per capita and lifetime value (LTV), as well as the system’s p99 latency and resource overhead for infrastructure, showed no significant change in our experiment.

Human Evaluation Human evaluation for user experience is a crucial step before deploying a new algorithm online. We have a strict in-house criterion that defines top-grade description. For experiment and control groups, we randomly sample 400 descriptions per group for evaluation. Results show that the generated descriptions of INTERACTOR achieve **+7pp** on top-grade rate, demonstrating that our policy achieves better user experience compared to previous production policy.

Beyond this, our ads system undergoes continuous, large-scale monthly human evaluation, where randomly-sampled impression creatives are collected to perform human evaluation for monitoring long-term quality. The INTERACTOR was deployed to serve the main traffic in the last week of May 2026, and the routine evaluation in June 2026 demonstrated a sustained quality improvement of **+9pp** on top-grade rate compared to the first three weeks in May 2026. This robustly proves that INTERACTOR improves true creative quality.

5 Conclusion

We identify the necessity to shift from generating CTR-driven ad titles to informative ad descriptions in industrial practice, highlighting the importance of explicitly modeling world knowledge to

satisfy user intent. We propose INTERACTOR, a multi-turn iteration framework optimized via agentic RL to achieve continuous improvement on multi-dimensional evaluations. Online A/B test verifies its effectiveness in improving ad revenue and user experience for an industry-scale system.

Ethical Considerations

When we apply AIGC technologies for online advertising business, both content security and user privacy protection are crucial. We make the following efforts to ensure only secure contents will be pushed online and alleviate potential privacy risks: 1) We only generate ad creatives for the advertisers who have submitted authorization to our advertising platform. 2) We have a strict risk-control process before deployment, which ensures the content security via a large number (over 100) of rules and high performing models to delete harmful results and rule-breaking contents, covering the recognition capacity for various trivial details about negative content. 3) Our algorithm does not incorporate any user privacy information (such as age and gender), which is not collected in the dataset.

To substantiate the safety of deploying INTERACTOR, we provide additional details about the risk-control process and its empirical performance, especially to sensitive categories.

We first report the faithfulness and risk-control deletion rates for sensitive categories. On the hold-out test set, the Average@10 faithfulness scores of INTERACTOR are 0.859 for the health category and 0.923 for the financial service category. During offline generation, the daily deletion rates by the risk-control process are 4.3% for the health category and 2.7% for the financial service category. These results characterize the quality of generated outputs and the behavior of the downstream risk-control process on sensitive categories.

We further report the effectiveness of the risk-control process for INTERACTOR outputs. Before the online experiment, we accumulated over 10k samples that were labeled as containing phrases unsuitable for display (e.g., including exaggerated statements inconsistent with the landing page) by human evaluation. We passed these samples through the risk-control process, and none of the samples remained, corresponding to an observed residual rate of 0%. This provides direct evidence that the risk-control process effectively identifies unsafe outputs represented in INTERACTOR.

⁷All offline generated creatives are fed into a risk-control process to delete unsecured results before being stored to KV database, and real-time generated titles are fed into an online risk-control stage before the ranking stage.

Regarding category-level restrictions, the risk-control algorithm applies stricter criteria to sensitive categories. For example, it prohibits potentially unsupported claims about treatment, diagnosis, or medication without sufficient evidence in health ads, as well as unsupported claims concerning investment, lending, or credit outcomes in financial ads. The risk-control rules and models are continuously updated based on new patterns of negative content identified through routine human evaluation and reports from advertisers and users. The same risk-control process covers both LLM-generated and advertiser-uploaded creatives.

We also maintain a real-time intervention pathway that can directly process reports from advertisers and users and remove unsafe content. On the advertiser side, for example, advertisers may submit unreasonable keywords to compete with competitors, which can trigger intervention. On the user side, an appeal button is provided at the ad display position for users to report problematic ads. Our system can process these reports and perform real-time intervention when necessary.

Therefore, the above safeguards in our system build a comprehensive protection mechanism: 1) strict criteria to all categories especially for sensitive ones, 2) dedicated risk-control rules and models applied to both generated and advertiser-uploaded creatives, 3) effective risk-control process covering each search request, and 4) post-deployment monitoring and real-time intervention.

Limitations

Our proposed framework focuses on automated ad description generation that provides high-quality candidates to advertisers, however the corpora of advertiser-written descriptions are not explicitly exploited during our framework development (they only implicitly contribute to the development of GenRMs). We believe that our environment possesses the ability of providing editing suggestion for them, and thus how to mine the value of human-written corpora and build a human-in-the-loop optimization paradigm for ad text generation is a promising direction for further research.

Acknowledgments

We thank all the anonymous reviewers for their valuable comments. We also thank Yi Zhang for his kind assistance.

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901.
- Yanda Chen, Zihui Ren, Qixiang Gao, Jiale Chen, Si Chen, Xubin Li, Tiezheng Ge, and Bo Zheng. 2025. CTR-driven ad text generation via online feedback preference optimization. *arXiv preprint arXiv:2507.20227*.
- DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Wenxin Dong, Chang Gao, Guanghui Yu, Xuewu Jiao, Mingqing Hu, Qiang Fu, Peng Xu, Penghui Wei, Hui Xu, Yue Xing, Shuanglong Li, and Lin Liu. 2026. Efficient LLM-based advertising via model compression and parallel verification. *arXiv preprint arXiv:2605.11582*.
- GLM. 2026. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M Hendryx. 2026. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text de-generation. In *International Conference on Learning Representations*.
- Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*.
- J Weston Hughes, Keng-hao Chang, and Ruofei Zhang. 2019. Generating better search engine text advertisements with deep reinforcement learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2269–2277.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*.
- Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. 2021. An empirical study of generating texts for search engine advertising. In *Proceedings of the 2021 Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, pages 255–262.
- Yashal Shakti Kanungo, Sumit Negi, and Aruna Rajan. 2021. Ad headline generation using self-critical masked language model. In *Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: human language technologies: industry papers*, pages 263–271.
- Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*.
- Kimi. 2025. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.
- MiniMax. 2026. The MiniMax-M2 series: Mini activations unleashing max real-world intelligence. *arXiv preprint arXiv:2605.26494*.
- Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. 2024. Striking gold in advertising: Standardization and exploration of ad text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 955–972.
- Qwen. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Tianxiao Shen, Myle Ott, Michael Auli, and Marc’Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International conference on machine learning*, pages 5719–5728. PMLR.
- Chang Wang, Siyu Yan, Depeng Yuan, Yuqi Chen, Yanhua Huang, Yuanhang Zheng, Shuhao Li, Yinqi Zhang, Kedi Chen, Mingrui Zhu, and Ruiwen Xu. 2026. DIVER: Unlocking diversity in ad headline generation with large language models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4981–4986.
- Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. 2021. Reinforcing pretrained models for generating attractive text advertisements. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3697–3707.
- Yongzhen Wang, Jian Wang, Heng Huang, Hongsong Li, and Xiaozhong Liu. 2020. Evolutionary product description generation: A dynamic fine-tuning approach leveraging user click behavior. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 119–128.
- Penghui Wei. 2025. IDEATION: Diversifying ad text generation in sponsored search with latent diffusion. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1394–1397.
- Penghui Wei, Xuanhua Yang, Shaoguo Liu, Liang Wang, and Bo Zheng. 2022. CREATER: CTR-driven advertising text generation with controlled pre-training and contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 9–17.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, Clark Barrett, and Ying Sheng. 2024. SGLang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Approach	Basic.	Info.	Faith.	Attract.
FEW-SHOT	0.985	0.521	0.564	7.1%
REACT	0.985	0.526	0.589	6.9%
REACT w/ Custom. Env.	0.984	0.631	0.683	6.7%
SFT w/ Best-of- N	0.986	0.685	0.711	11.6%

Table 4: Additional results on off-the-shelf LLMs under few-shot and test-time scaling settings, as well as the SFT checkpoint.

A Additional Results on Methods with Off-the-shelf LLMs

During early experiments, we also evaluate frontier LLMs under few-shot and test-time scaling settings to explore the potential of multi-turn iteration.

- FEW-SHOT (Brown et al., 2020): We prompt the Gemini-3.1-Pro⁸ with detailed rubrics of multi-dimensional quality requirements and few-shot examples to generate descriptions.
- REACT (Yao et al., 2023): We prompt the Gemini-3.1-Pro to iteratively critique and improve its own generations, without the explicit reward signals used in INTERACTOR.
- REACT w/ Customized Env.: It further incorporates the search engine and GenRMs used in the environment of INTERACTOR to provide knowledge and feedback during iteration.

We also compare the test-time scaling settings of the SFT checkpoint:

- SFT w/ Best-of- N : It first generates $N = 20$ candidates with the SFT checkpoint of SINGLE-TURN RL, and then uses our reward signal to select 10 best-performing outputs.

Table 4 verifies that INTERACTOR outperforms these baselines by a large margin, especially on attractiveness which reflects user dynamic preference. User behaviors in search ads are highly dynamic and time-varying, and an LLM relying on static few-shot prompts cannot capture such dynamics and lacks alignment. The improvement on informativeness and faithfulness through incorporating customized environment further demonstrates the necessity of constructing accurate reward signals based on quality requirements.

⁸<https://deepmind.google/models/model-cards/gemini-3-1-pro/>

B Implementation Details

B.1 Hybrid Rewards in the Environment

We build the GenRMs of knowledge capacity and landing page consistency to transform open-ended evaluation to verifiable signals. We first summarize the rubrics and then prompt Qwen3-30B-A3B (Qwen, 2025) to obtain binary result of the reward and fine-grained feedback as reasoning information. Table 5 lists the predefined rubrics and prompts of the GenRMs (without few-shot examples for brevity). The prompt is optimized via a continual check of the consistency between the GenRM and human annotators, and finally the consistency rate is above 0.8 on 200 randomly sampled examples, verifying their usability.

For the CTR estimation model, we use the Qwen3-Embedding-0.6B (Zhang et al., 2025) as base model with an extra classification head to discriminate click or unclick,⁹ where the input contains both x_{user} , x_{ad} and a generated description. We then perform continual training on the user click logs of our search ads system. The AUC on test set is above 0.7 and the PCOC (predicted CTR over the true CTR) is 1.04, verifying its usability as an offline model.

B.2 Evaluation Metrics

To evaluate basic quality, informativeness and faithfulness of generated descriptions, we employ the LLM-as-a-Judge method (Zheng et al., 2023) via prompting DeepSeek-V3.2 (DeepSeek-AI, 2025) to return binary results. For basic quality, Table 6 shows the prompt. For informativeness and faithfulness, we use the same prompts of the GenRMs for knowledge capacity and landing page consistency to perform evaluation, as shown in Table 5.

To evaluate attractiveness, we train a larger model based on Qwen3-Embedding-8B, using the same procedure as the CTR estimation model in policy learning.

To evaluate the diversity of multiple generations, we use $1 - \text{PairwiseBLEU}$ (Shen et al., 2019) that measures the dissimilarity across K candidates $\{y_i\}_{i=1:K}$ per input ($K = 10$ in our experiment):

$$\text{PairwiseBLEU} = \frac{1}{K(K-1)} \sum_{i \neq j} \text{BLEU}(y_i, y_j). \quad (4)$$

⁹We do not employ the token probability of ‘yes’ or ‘no’ because a pointwise CTR prediction model should meet the calibration condition, which produces probabilistic predictions reflecting true likelihoods.

B.3 Policy Learning and Inference

We use Slime for training with $8 \times$ NVIDIA H800 GPUs, and the base model is Qwen3-30B-A3B. For the environment, we deploy a remote server for the reward models, and request our in-house search engine with internal API. The maximum turn is set to 3, and the group size is 8. The factors of multiple rewards are set to 1, 2 and 5. We configure fully asynchronous to acceleration and employ routing replay (Zheng et al., 2025) to stabilize MoE training. For fair comparison, all competitors use the same set of 100k synthesized descriptions by DeepSeek-V3.2 for cold-start SFT. Table 7 lists the task prompt of INTERACTOR.

We use SGLang (Zheng et al., 2024) for efficient offline batching inference. For each input we generate multiple candidates with top- p sampling (Holtzman et al., 2020). Although it seems that INTERACTOR increases inference cost compared to the baseline, its computational resource utilization has been improved in industrial deployment. In the baseline deployment, descriptions are first generated and then filtered by post-processing models, therefore the compute spent on the discarded results is wasted. The INTERACTOR essentially transfers post-processing into multi-turn iteration and eliminates extra post-processing, ultimately achieving more qualified results.

C Related Work

Recent ad text generation approaches have gradually evolved from small models (Hughes et al., 2019; Wang et al., 2020; Kamigaito et al., 2021; Kanungo et al., 2021; Wang et al., 2021; Wei et al., 2022) to large models (Mita et al., 2024; Chen et al., 2025; Wang et al., 2026; Wei, 2025) with preference optimization (Rafailov et al., 2023) or RLHF (Ziegler et al., 2019) for improvement.

As the pioneering work, Chen et al. (2025) pre-collected static preference data to achieve higher CTR, and Wang et al. (2026) proposed to jointly optimize diversity, quality and CTR via multi-objective RLHF. However, current work focuses on ad title, while approaches for high-quality ad description remain underexplored. We point out the necessity to shift to knowledge-rich ad description, and inspired by agentic systems (Yao et al., 2023) and agentic RL paradigm (Jin et al., 2025; GLM, 2026; MiniMax, 2026; Kimi, 2025) we propose a practical framework via multi-turn iteration in industrial-scale search ads system.

Quality Dimension	Prompt
Knowledge Capacity	### Evaluation Criteria Given the ad landing page information, the retrieved content via searching web pages based on the user query, and the candidate description, please evaluate strictly based on the following rules and provide a judgment of "pass" or "fail". ### Dimension: Knowledge Capacity - The definition of world knowledge is: any knowledge that meets any of the following conditions: - Explain professional knowledge such as definitions, backgrounds, principles, and concepts (e.g., introduce "what is xx", "why is xx", "what are the symptoms of xx disease", etc.). - Describe the characteristics / attributes / applicable scenarios / mechanism of action / working principle / efficacy, etc. of things. - Popularize the objective impact of objects/behaviors. - Provide a neutral reference (e.g., "Usually, solutions to such problems include categories A, B, and C"). ### Execution steps Step 1: Confirm whether the candidate description contains knowledge from the retrieved content Step 2: If there is knowledge from the retrieved content, determine: Rule 2.1: The contained knowledge meets the definition of world knowledge Rule 2.2: The knowledge is consistent with the services that can be provided on the landing page The knowledge capacity is considered to be passed only when both of the above rules are met Step 3: If there is no knowledge from the retrieved content, determine: Rule 3.1: the retrieved content is null Rule 3.2: Inconsistency between the brand/platform/software/app/website/organization/company mentioned in the retrieved content and the landing page As long as either of the above two conditions is met, the level of knowledge is considered as passed ### Output format Structured output format: Output strictly in accordance with the following JSON format, without any additional text: <pre>{ "assessment": { "knowledge_capacity": { "verdict": "Pass/Fail", "reason": "First, determine whether there is knowledge from the retrieved content in the candidate description. Then, based on this, briefly provide specific reasons for judging whether the degree of knowledge capacity has been achieved." } } }</pre>
Landing Page Consistency	### Task Description You are a professional ad description evaluation expert, skilled in rigorously assessing the content quality. Your task is to conduct an assessment of [Landing Page Consistency] based on all the rules, and provide a conclusion of "Pass" or "Fail". ### Evaluation Criteria Check whether the content in the candidate ad description can be supported by the provided landing page information. The criteria of Pass: the following rules must be met simultaneously. Functional authenticity: All specific functions, services, data, effects, offers, prices, parameters, and time-sensitive commitments mentioned in the description (such as "free trial", "efficiency improvement by 50%", "buy one get one free", "24-hour customer service") must be found with completely identical or equivalent expressions in landing page. Brand consistency: All brand, company, organization, website, and App names mentioned in the description (whether owned by the advertiser or its partners) must appear in landing page. ### The evaluation steps you must follow - Deconstruction: Carefully read the Candidate Description and list each item one by one: All function points (Function Point 1, Function Point 2, Function Point 3...) All brand mentions (Brand 1, Brand 2, Brand 3...) - Validation: - For each feature, search and verify in the landing page. Record the basis found, or mark "not found". - For each brand, search and verify in the landing page. Record the basis found, or mark "not found". - Adjudication: - If all functional points have been justified, then 'function_verdict' = "Pass"; otherwise, 'function_verdict' = "Fail". - If all brands have been justified, then 'brand_verdict' = "Pass"; otherwise, 'brand_verdict' = "Fail". - Conclusion: The final conclusion is "Pass" only when both 'function_verdict' and 'brand_verdict' are "Pass". ### Important Principles No Assumptions: Even if you believe a certain feature is reasonable or common, as long as you cannot find a clear basis in the provided landing page, you must determine it as "Fail". No external knowledge allowed: Your judgment must be 100% based on the input information given above. Synonym judgment: For functional points, different expressions are allowed, but the core meaning must be exactly the same (e.g., "3-year warranty" and "warranty period of 36 months" can be considered synonymous). ### Output format Structured Output Format: Output strictly in accordance with the following JSON format, without any additional text: <pre>{ "assessment": { "landing_page_consistency": { "function_verdict": "Pass/Fail", "function_reason": "List how you conducted comparative judgments on the functional level, what evidence you found, or which functional point you could not find evidence for.", "brand_verdict": "Pass/Fail", "brand_reason": "List how you conducted comparative judgments on the brand level, what evidence you found, or which brand you mentioned but could not find evidence for." } } }</pre>

Table 5: Predefined rubrics and prompts for the GenRMs. Translated to English.

Quality Dimension	Prompt
Basic Quality	### Task Description You are a professional ad description evaluation expert, skilled in rigorously assessing the content quality. Your task is to evaluate the [basic quality] based on all the rules, and provide a conclusion of either "Pass" or "Fail". ### Evaluation Criteria Check whether the content in the candidate ad description meets the standards, with a focus on the following common language issues: 1. Redundancy: Words, phrases, or sentences that are redundant (e.g., "Limited time offer, offer for a limited time"). 2. Semantic incompleteness: a. The absence of internal components in the sentence leads to unclear or unintelligible meaning expression b. The sentence ends with "." or "..." 3. Mixed sentence patterns: The sentence is internally complete, but the sentence patterns are mixed together among multiple sentences, resulting in a chaotic structure and incoherent flow 4. Punctuation errors: a. Missing punctuation b. Misuse of punctuation c. Improper use of punctuation (such as meaningless spaces) 5. Undecipherable characters/tipos: a. Unrecognized characters, symbols, or encoding errors appear in the text (such as "click to view &gt;&gt;") b. Obvious tipos (using Chinese characters that look similar but have different meanings, or using words with the same pronunciation but different characters, e.g., "三日游" -> "三曰游") c. Professional terminology is altered by adding or changing characters (e.g., "银屑病" -> "银皮病") d. The sentence contains a file format name, such as ".doc", ".txt", etc e. Insert English letters into Chinese characters without forming English words or conveying special meanings f. The sentence contains content with website meanings such as "www.", ".cn", ".tv", etc ### The criteria of Pass: The basic quality is "Pass" only when none of the aforementioned five issues exists. Otherwise, it is "Fail". ### Evaluation steps you must follow 1. Reading: • Carefully read the candidate ad description 2. Check item by item: • For each issue (repetition, incomplete semantics, mixed sentence structures, punctuation errors, garbled text/tipos), determine whether it exists and record the basis for the determination • If any other types of linguistic issues are found, please specify them in 'final_verdict_reason' 3. Adjudication: • Based on the inspection results, provide the "Pass/Fail" status and reasons for each sub-dimension • Based on all sub-dimensions, a final conclusion 'final_verdict' is drawn: if all sub-dimensions are "Pass", then 'final_verdict' = "Pass"; otherwise, it is "Fail" ### Output format Structured Output Format: Output strictly in accordance with the following JSON format, without any additional text: <pre>{ "assessment": { "content_quality": { "final_verdict": "Pass/Fail", "final_verdict_reason": "If it fails, list the existing language quality issues." } } }</pre>

Table 6: Predefined rubrics and prompt for the evaluation of basic quality. Translated to English.

Task Prompt
Based on the given inputs of user search query and ad landing page information, generating a high-quality ad description that integrates appropriate world knowledge to satisfy user intent and contains faithful selling points from landing page to represent product advantages, through a multi-turn iteration process: integrating retrieved knowledge relevant to query, requesting reward signals for obtaining fine-grained feedbacks of initial generation, and iteratively refining it with the received feedbacks. Execution steps: Think: Every time you receive new information, you must first conduct reasoning between <think> and </think>. Retrieval Knowledge: At the beginning of execution, if you find that you lack relevant information for the user query, you can call the search engine using <retrieve> Retrieve relevant knowledge of [user query] </retrieve>, and the retrieved content will be returned between <information> and </information>. Generate: Place the generated description between <create> and </create>, immediately followed by <reward> request reward models </reward> for requesting rewards that providing detailed feedbacks about knowledge capacity and landing page consistency. Iteratively Refine: After receiving the fine-grained feedback enclosed in <check> and </check>, please optimize the generated description based on the feedback.

Table 7: The task prompt of our INTERACTOR framework. Translated to English.